



Une structure compacte pour indexer un fichier ordonne et son evaluation

D. Plateau

► To cite this version:

D. Plateau. Une structure compacte pour indexer un fichier ordonne et son evaluation. RR-0169, INRIA. 1982. inria-00076389

HAL Id: inria-00076389

<https://hal.inria.fr/inria-00076389>

Submitted on 24 May 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



CENTRE DE ROCQUENCOURT

Rapports de Recherche

N° 169

UNE STRUCTURE COMPACTE POUR INDEXER UN FICHER ORDONNÉ ET SON ÉVALUATION

Didier PLATEAU

Novembre 1982

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
BP 105
78153 Le Chesnay Cedex
France
Tel. 954 90 20

UNE STRUCTURE COMPACTE POUR INDEXER UN FICHIER ORDONNE ET SON EVALUATION

Didier PLATEAU

Résumé

L'un des critères majeurs de performance d'un système de gestion de base de données est la vitesse d'accès aux données. Cette vitesse dépend de nombreux facteurs (technologie, organisation de l'espace mémoire, chemins d'accès...). Parmi ces facteurs nous nous intéressons au problème de l'indexation d'un fichier totalement ordonné. Une importante amélioration des performances du système peut être obtenue si l'index est maintenu résident en mémoire centrale. Dans ce but, nous présentons une structure de type arbre lexicographique (TRIE cf. [Knuth]), dont la propriété est d'être très compacte. Une analyse de la taille moyenne est proposée sous diverses hypothèses. En nous replaçant dans le cadre particulier de la machine base de données VERSO, nous utilisons les résultats de l'analyse pour montrer que nous pouvons maintenir l'index résident et le manipuler de façon performante.

Abstract

One of the main performance criteria in a data base management system is data access speed. This speed depends on many factors such as technology, memory organization, access paths ... Among those problems, we are interested in indexing a sorted file. An important improvement of the system performance is expected if the index is resident in main memory. We present a TRIE structure [Knuth], the main feature of which is compactness. An analysis of this structure average size is proposed, under different assumptions. Placing this work in the framework of the data base machine VERSO [VERSO], we show that we can maintain the index resident in main memory.



1. INTRODUCTION

L'accès aux données constitue l'un des problèmes majeurs dans un Système de Gestion de Base de Données (SGBD). Sa vitesse met en jeu différents facteurs : technologie, organisation de la mémoire, chemins d'accès...

Nous nous intéressons plus spécialement au problème de l'indexation d'un fichier totalement ordonné. Afin de maintenir de bonnes performances, la plupart des SGBD utilisent un mécanisme d'indexation des données qui est utilisé une ou plusieurs fois à chaque accès aux données. Il est donc important que ces manipulations soient rapides.

Pour réaliser cet index, deux types de solutions sont généralement adoptées : hachage, ou séquentiel indexé.

La propriété d'une fonction de hachage est de briser l'ordre inhérent aux unités logiques de données, et de permettre la localisation rapide d'une unité.

Pour indexer un fichier ordonné et répondre à des requêtes mettant en jeu des notions d'ordre, cette méthode n'est pas appropriée.

Par ailleurs, les solutions classiques de séquentiel indexé (listes inversées, Arbres B), bien adaptées à la notion d'ordre, n'offrent pas une compaction suffisante pour que l'index soit maintenu résident en mémoire centrale.

Nous allons décrire une structure de type arbre lexicographique (ou TRIE) se rattachant aux méthodes de séquentiel indexé et présentant la propriété d'être très compacte. Notons qu'une approche très similaire a déjà été proposée dans [Litwin] sous le nom de TRIE HASHING. La section 1 fournit une description précise du contexte dans lequel nous nous plaçons, et de la structure choisie. L'utilisation de cette structure, pour une "application raisonnable", permettra de maintenir l'index résident

en mémoire centrale et donc le manipuler rapidement. Nous proposons une évaluation de la taille moyenne de l' index. Plusieurs valeurs sont obtenues, dépendant des hypothèses faites sur la nature du fichier indexé (distribution des valeurs des unités de données, indépendance de ces valeurs), et aussi des approximations faites pour les obtenir (longueur fixe, longueur infinie ...). Les techniques utilisées sont de type série génératrices et transformation de Mellin, dont une présentation est faite dans [Flajolet, Sotteau].

En nous plaçant dans le contexte de la machine base de données VERSO [VERSO], nous montrons que cette structure permet de maintenir l'index résident .

2. DESCRIPTION DES DONNEES ET DE LA STRUCTURE DE L'INDEX

On suppose ici le fichier totalement ordonné. Le rôle d'un index est de créer un lien entre l'organisation des données au niveau logique et au niveau physique.

Nous définissons donc le modèle logique des données, leur représentation physique, et la structure de l'index.

Le modèle logique des données est relationnel. (Nous supposerons le lecteur familier avec les concepts relationnels). La structure logique des données est donc la table :

Exemple : Soit la relation $NE(E) \leftarrow A[VILLE, NOM]$;

Une valeur possible de $NE(E) \leftarrow A$ est :

VILLE	NOM
AUBENAS	FRANCOIS
PARIS	MICHEL
AMSTERDAM	CHRISTIAN
AMIENS	PHILIPPE
PARIS	ANNE
PARIS	SOPHIE
PARIS	ANDRE
ANVERS	SERGE
ANTIBES	JEAN
PARIS	ANTOINE

Figure 1. Représentation en table d'une relation

Physiquement, les données sont stockées sur une mémoire secondaire dont l'unité adressable est le bloc. Chaque bloc contient un nombre entier (a priori non constant) de n-uplets.

Les attributs sont des chaînes de caractères (ou d'octets).

En considérant le n-uplet comme une chaîne de caractères obtenue par concaténation de ses attributs, on peut alors munir l'ensemble des n-uplets de l'ordre lexicographique.

A la relation est associée une liste de blocs B_1, \dots, B_k telle que :

$\forall i < j$ tous les n-uplets de B_i sont inférieurs aux n-uplets de B_j

Exemple : à la relation NE(E)-A définie figure 1, on associe la liste de blocs :

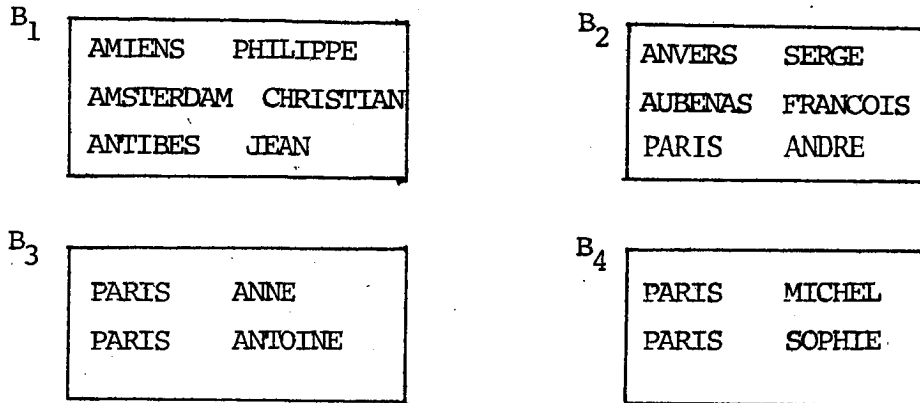


Figure 2. Représentation physique d'une relation

Structure de l'index :

Définissons tout d'abord la notion d'arbre lexicographique (AL) et d'arbre lexicographique compacté (ALC).

Définition :

Soit Σ un alphabet ordonné ; $|\Sigma| = \alpha$;

Un arbre lexicographique sur Σ est un arbre de degré α , dont chaque sommet est étiqueté par un caractère de Σ ; à chaque sommet correspond une séquence de sous-arbres ordonnée totalement par l'étiquette de leur sommet.

Un arbre lexicographie compacté est un arbre lexicographique dans lequel on a compacté les branches filiformes terminales en une seule feuille, elle-même étiquetée par l'étiquette du sommet de la branche.

Exemple : Soit $\Sigma = \{A, \dots, Z, \#\}$
un exemple d'AL est :

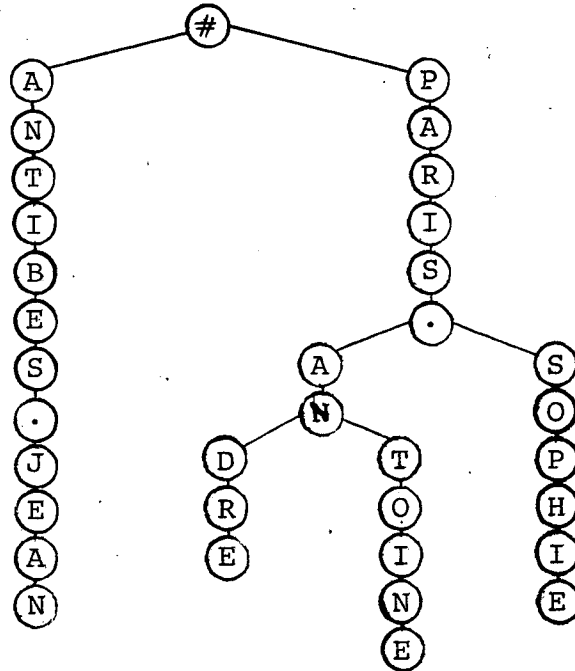


Figure 3. Exemple d'arbre lexicographique

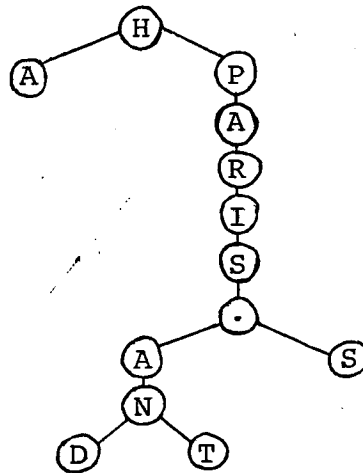


Figure 4. L'ALC associé à l'AL défini figure 3.

A chaque feuille d'un AL, on fait correspondre la chaîne formée par concaténation des étiquettes suivies lorsqu'on parcourt l'arbre du sommet vers la feuille.

A un arbre, on associe donc l'ensemble des chaînes de caractères associées à ses feuilles ; et réciproquement à un ensemble de chaînes de caractères, on associe cet AL, puis l'ALC obtenu en compactant.

Exemple : On met en correspondance les arbres des figures 3 et 4 avec : {AMIENS,AMSTERDAM,ANTIBES,ANVERS,AUBENAS}

Ces définitions données, nous pouvons décrire la structure logique de l'index associé à un fichier : l'index est l'ALC associé à l'ensemble des chaînes de caractères représentant les plus grands n-uplets de chaque bloc de la relation.

Ces chaînes sont précédées d'un caractère réservé "#". En plus de son étiquette, chaque feuille contiendra l'adresse du bloc correspondant.

Exemple. à la relation définie figures 1 et 2, on associe l'index :

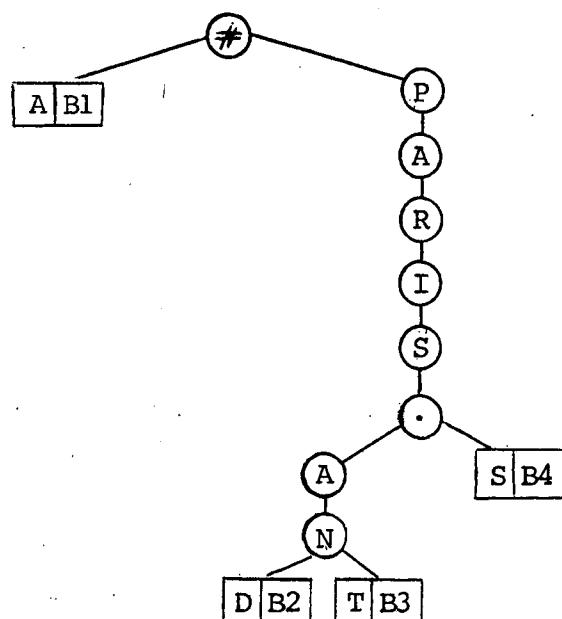


Figure 5. l'index sous forme d'ALC.

Physiquement, l'index est obtenu en linéarisant l'AIC par parcours de l'arbre en ordre préfixe. Une parenthèse est ouverte à chaque passage à un sommet ayant plus d'un fils, puis refermée lorsqu'on revient à son niveau. Le caractère "(", signale que l'on est au niveau d'une feuille.

Exemple. l'arbre présenté figure 5 a la représentation linéaire :

(A (B1 PARIS. (AN (D (B2 T (B3) S (B4))

C'est désormais à la représentation physique de l'index que nous nous intéressons.

3. EVALUATION DE LA TAILLE DE L'INDEX

Pour calculer la taille moyenne de l'index, il est nécessaire de définir un modèle. Ce modèle inclut des hypothèses probabilistes sur la distribution des valeurs des n-uplets de l'index et des paramètres (taille du fichier, taille moyenne des n-uplets ...).

Nous proposons deux modèles qui diffèrent par leurs hypothèses probabilistes.

Modèle 1 : Uniformité et indépendance globale

Les paramètres du modèle sont la longueur moyenne des n-uplets de la relation ℓ et le nombre de blocs occupés par la relation B. (Notons que B est aussi le nombre de n-uplets de l'index).

Nous supposons, pour l'évaluation de la taille de l'index, que les n-uplets sont de longueur fixe ℓ et qu'ils prennent leurs valeurs uniformément dans un langage Σ^ℓ , indépendamment des autres n-uplets.

Notons $\overline{SOC}(\ell, B)$ le nombre moyen de sommets internes d'un ALC représentant B chaînes de caractères de longueur ℓ , sous hypothèse d'uniformité,

et $\overline{SOC1}(\ell, B)$ le nombre moyen de sommets internes d'un ALC représentant B chaînes de caractères de longueur ℓ , tels que ces noeuds n'ont qu'un seul fils, sous hypothèse d'uniformité.

Alors la taille $T(\ell, B)$, le nombre d'octets, d'un index adressant B blocs d'une relation dont les n-uplets sont de longueur moyenne ℓ est donnée, dans ce modèle 1, par la formule :

$$T(\ell, B) = \overline{SOC1}(\ell, B) * c + (\overline{SOC}(\ell, B) - \overline{SOC1}(\ell, B)) * 3c \\ + B * (2c + b)$$

où c est la taille, en nombre d'octets, d'un caractère et b est la taille d'une adresse de blocs.

En effet, à chaque sommet interne n'ayant qu'un fils correspond l'étiquette du sommet ; à chaque sommet interne ayant plus d'un fils correspondent l'étiquette du sommet, et deux parenthèses ; enfin, à chaque feuille correspondent l'étiquette de la feuille, le caractère réservé "□", et l'adresse du bloc.

Modèle 2 : Uniformité et indépendance des attributs

Soit Σ un alphabet ($|\Sigma| = \alpha$), $R[A_1, \dots, A_n]$ une relation. On suppose que chaque A_i est une chaîne de caractères de Σ , de longueur moyenne ℓ_i ; la relation occupe B blocs.

On note $S_i = \frac{|\tilde{R}[A_1 \dots A_i]|}{B}$ $1 \leq i \leq n$ qu'on appellera la sélectivité de $A_1 \dots A_i$.

$|R[A_1 \dots A_i]|$ note le nombre de valeurs distinctes de $A_1 \dots A_i$ apparaissant dans l'index.

Pour effectuer l'évaluation de la taille de l'index, on suppose que chaque A_i est de taille fixe ℓ_i , et prend des valeurs uniformément dans Σ^{ℓ_i} , indépendamment des autres valeurs de A_i , et des valeurs des autres attributs du n-uplet.

Par exemple, dans le cas $S_1=1$, chaque valeur de l'attribut A_1 détermine (aux sens que le relationnel donne à ce verbe) le n-uplet. La valeur de A_1 suffit donc à discriminer les n-uplets de l'index. Du fait de la structure d'ALC, seules les valeurs de A_1 apparaissent dans l'index

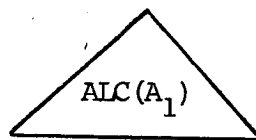


Figure 6. Schéma de l'index dans le cas $S_1=1$

Par contre, si S_1 est très petit et si S_2 vaut 1, à chaque valeur de A_1 correspond un grand nombre de n-uplets tandis qu'une valeur de $A_1 A_2$ détermine le n-uplet. Dans ce cas l'index sera composé d'un AL représentant les $B \cdot S_1$ (en moyenne) valeurs de A_1 , et de $B \cdot S_1$ ALC représentant chacun en moyenne les S_1^{-1} valeurs de A_2 associées à chaque valeur de A_1 .

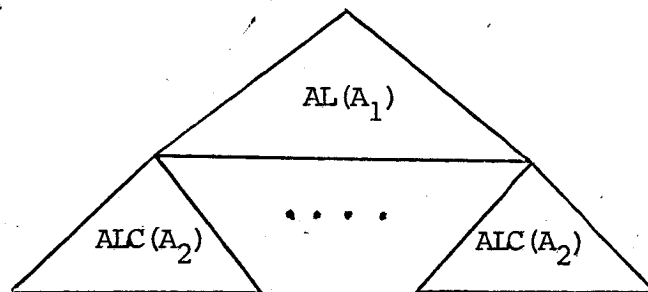


Figure 7. Schéma de l'index dans le cas S_1 très petit et $S_2=1$

Ceci n'est bien sûr qu'une approximation dans la mesure où nous raisonnons en terme d'espérances.

Dans le cas général, notons i le plus petit indice tel que $S_i=1$; (un tel i existe car $S_1 \leq \dots \leq S_n = 1$). Alors l'index aura l'allure suivante :

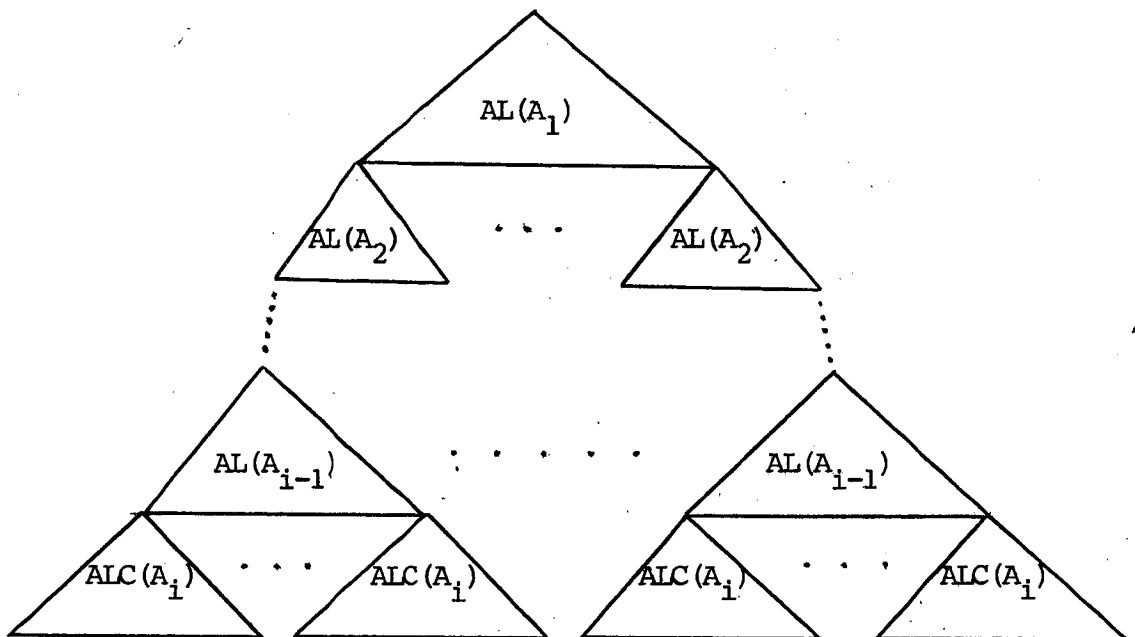


Figure 8. Schéma de l'index quand i est le plus petit indice tel que $S_i = 1$;

Notons :

- * $\overline{SOC}(\ell, B), \overline{SOCI}(\ell, B)$ les paramètres définis dans le modèle 1 ;
- * $\overline{SOM}(\ell, B)$ le nombre moyen de sommets d'un AL représentant B chaînes de caractères de longueur fixe ℓ , sous hypothèse d'uniformité
- * $\overline{SOMI}(\ell, B)$ le nombre moyen de sommets n'ayant qu'un fils, d'un AL représentant B chaînes de caractères de longueur fixe ℓ , sous hypothèse d'uniformité

Alors la taille moyenne d'un index représentant B n-uplets composés chacun de n attributs A_1, \dots, A_n (auxquels on associe les paramètres ℓ_1, \dots, ℓ_n et S_1, \dots, S_n comme définis précédemment) sera $\varphi(1, B)$ où la fonction φ est définie au point (i, x) par

$$\text{Si } S_i = 1 \text{ alors } \varphi(i, x) \leftarrow \overline{\text{SOC}} \, 1(\ell_i, x) * c + (\overline{\text{SOC}}(\ell_i, x) - \overline{\text{SOC}} \, 1(\ell_i, x)) * 3c \\ + x * (2c + b) ;$$

$$\text{Sinon } \varphi(i, x) \leftarrow \overline{\text{SOM}} \, 1(\ell_i, x * S_i) * c + (\overline{\text{SOM}}(\ell_i, x * S_i) - \overline{\text{SOM}} \, 1(\ell_i, x * S_i)) * 3c \\ + x * S_i * \varphi(i+1, S_i^{-1}) ;$$

L'avantage du deuxième modèle est de prendre en compte la notion de sélectivité des attributs. Dans un exemple comme la relation NE(E)-A, il est clair que chaque VILLE va apparaître un grand nombre de fois et que l'hypothèse d'uniformité globale n'est pas réaliste.

Néanmoins, la complexité du deuxième modèle, peut provoquer une accumulation dangereuse d'approximations : un raisonnement sur les espérances est bon si la variance des paramètres est faible mais mauvais dans le cas contraire. Nous comparerons les résultats des deux modèles, lors des applications numériques.

Nous allons maintenant évaluer les paramètres $\overline{\text{SOC}}(\ell, B)$, $\overline{\text{SOC}} \, 1(\ell, B)$, $\overline{\text{SOM}}(\ell, B)$ et $\overline{\text{SOM}} \, 1(\ell, B)$. Pour chacun d'eux, nous proposons :

- (1) une valeur exacte sous hypothèse de longueur fixe, (celle obtenue en calculant la série génératrice associée au paramètre).
- (2) une valeur approchée sous hypothèse de longueur fixe obtenue en manipulant l'expression (1).
- (3) une valeur approchée sous hypothèse de longueur infinie obtenue en calculant la série génératrice exponentielle associée au paramètre puis en utilisant la transformation de Mellin pour obtenir des valeurs plus compactes des expressions obtenues.

L'essentiel de ces techniques est décrit dans [Flajolet, Sotteau]

Les calculs de $\overline{\text{SOC}}(\ell, B)$ sont détaillés. Du fait de leur similitude, nous ne présentons que les étapes principales, de l'évaluation des trois autres paramètres.

EVALUATION DU NOMBRE MOYEN DE NOEUDS INTERNES D'UN ALC

Avant d'évaluer ce paramètre, nous allons reformuler le problème de façon générale et définir quelques notations ;

Soit * $\Sigma = \{1, \dots, \alpha\}$ un alphabet ;

* $\ell \in \mathbb{N}$; (\mathbb{N} est l'ensemble des nombres entiers positifs)

* $\text{SOC} : \mathcal{P}(\Sigma^\ell) \rightarrow \mathbb{N}$ qui, à toute partie p de Σ^ℓ , associe le nombre de noeuds internes de l'arbre lexicographique compacté représentant p .

(Σ^ℓ est l'ensemble des chaînes de ℓ caractères appartenant à Σ ; $\mathcal{P}(\Sigma^\ell) = \mathcal{P}^{(\ell)}$ est l'ensemble des parties de Σ^ℓ)

* $\mathcal{P}_n^{(\ell)}$ l'ensemble des parties à n éléments de $\mathcal{P}^{(\ell)}$

* $\mathcal{P}^{(\ell)}(x) = \sum_{p \in \mathcal{P}^{(\ell)}} x^{|p|}$;

* $\text{SOC}^{(\ell)}(x) = \sum_{p \in \mathcal{P}^{(\ell)}} \text{SOC}(p) \cdot x^{|p|}$;

* Pour $p \in \mathcal{P}^{(\ell)}$, on note p_j l'ensemble des chaînes de p commençant par le caractère j , privées de ce caractère.

Notons que p_j est un élément de $\mathcal{P}^{(\ell-1)}$;

Notons que $\mathcal{P}^{(\ell)}(x)$ et $\text{SOC}^{(\ell)}(x)$ peuvent également s'écrire

$$\mathcal{P}^{(\ell)}(x) = \sum_{n=0}^{\alpha^\ell} p(n, \ell) x^n ;$$

$$\text{SOC}^{(\ell)}(x) = \sum_{n=0}^{\alpha^\ell} \text{SOC}(n, \ell) x^n ;$$

$p_n^{(\ell)}$ représente le nombre de parties à n éléments de $\mathcal{P}^{(\ell)}$ et vaut donc $\binom{\alpha^\ell}{n}$;

$$\text{SOC}(n, \ell) = \sum_{p \in \mathcal{P}_n^{(\ell)}} \text{SOC}(p) ;$$

Sous hypothèses d'uniformité, on a donc :

$$\overline{\text{SOC}}(n, \ell) = \frac{\text{SOC}(n, \ell)}{p(n, \ell)} ;$$

Pour obtenir $\overline{\text{SOC}}(n, \ell)$, nous allons d'abord calculer $\text{SOC}^{(\ell)}(x)$;
Soit $p \in \mathcal{P}^{(\ell)}$; la définition de l'AIC représentant p permet d'établir

$$\text{SOC}(p) = \mathbb{1}_{\{|p| \geq 2\}} + \sum_{i=1}^{\alpha} \text{SOC}(p_i) ; \quad (o)$$

$$\begin{aligned} \text{où } \mathbb{1}_{\{|p| \geq 2\}} &= 1 \text{ si } |p| \geq 2 ; \\ &= 0 \text{ sinon ;} \end{aligned}$$

On en déduit que :

$$\begin{aligned} \text{SOC}^{(\ell)}(x) &= \sum_{p \in \mathcal{P}^{(\ell)}} \text{SOC}(p) \cdot x^{|p|} \\ &= \sum_{p \in \mathcal{P}^{(\ell)}} \left(\mathbb{1}_{\{|p| \geq 2\}} + \sum_{i=1}^{\alpha} \text{SOC}(p_i) \right) x^{|p|} \\ &= \sum_{\substack{p \in \mathcal{P}^{(\ell)} \\ |p| \geq 2}} x^{|p|} + \sum_{p \in \mathcal{P}^{(\ell)}} \left(\sum_{i=1}^{\alpha} \text{SOC}(p_i) \right) x^{|p|} \\ &= \mathcal{P}^{(\ell)}(x) - 1 - \alpha^\ell x + \sum_{p_1, \dots, p_\alpha \in \mathcal{P}^{(\ell-1)}} \text{SOC}(p_i) \cdot x^{\sum_{i=1}^{\alpha} |p_i|} \end{aligned}$$

La transformation du 1er terme de la somme découle de la définition de $\mathcal{P}^{(\ell)}(x)$.

Celle du 2ème terme vient de l'hypothèse d'uniformité et d'indépendance : quand p décrit $\mathcal{P}^{(\ell)}$ alors p_1, \dots, p_α décrivent $\mathcal{P}^{(\ell-1)}$ indépendamment les uns des autres).

D'où

$$\text{SOC}^{(\ell)}(x) = \mathcal{P}^{(\ell)}(x) - 1 - \alpha^\ell x + \alpha \sum_{p_1, \dots, p_\alpha \in \mathcal{P}^{(\ell-1)}} \text{SOC}(p_1) x^{\sum_{i=1}^{\alpha} |p_i|}$$

(cette transformation vient de ce que l'hypothèse d'uniformité donne à tous les p_i un rôle identique).

$$\text{SOC}^{(\ell)}(x) = \mathcal{P}^{(\ell)}(x) - 1 - \alpha^\ell x + \alpha \sum_{p_2, \dots, p_\alpha \in \mathcal{P}^{(\ell-1)}} x^{|p_2|} \dots x^{|p_\alpha|} \left(\sum_{p_1 \in \mathcal{P}^{(\ell-1)}} \text{SOC}(p_1) x^{|p_1|} \right)$$

$$\text{SOC}^{(\ell)}(x) = \mathcal{P}^{(\ell)}(x) - 1 - \alpha^\ell x + \alpha \left(\mathcal{P}^{(\ell-1)}(x) \right)^{\alpha-1} \text{SOC}^{(\ell-1)}(x) \quad (1)$$

$$\frac{\text{SOC}^{(\ell)}(x)}{\mathcal{P}^{(\ell)}(x)} = 1 - \frac{1 + \alpha^\ell x}{\mathcal{P}^{(\ell)}(x)} + \alpha \frac{\text{SOC}^{(\ell-1)}(x)}{\mathcal{P}^{(\ell-1)}(x)}$$

(car, $\mathcal{P}^{(\ell)}(x) = (\mathcal{P}^{(\ell-1)}(x))^\alpha$).

Par récurrence, on tire :

$$\frac{\text{SOC}^{(\ell)}(x)}{\mathcal{P}^{(\ell)}(x)} = \sum_{i=0}^{\ell-1} \alpha^i \left(1 - \frac{1 + \alpha^{\ell-i} x}{\mathcal{P}^{(\ell-i)}(x)} \right) + \frac{\text{SOC}^{(0)}(x)}{\mathcal{P}^{(0)}(x)}$$

Comme $\text{SOC}^{(0)}(x) = 0$, en remplaçant $\mathcal{P}^{(\ell-i)}(x)$ par sa valeur, on obtient :

$$\text{SOC}^{(\ell)}(x) = \frac{\alpha^{\ell-1}}{\alpha-1} (1+x)^{\alpha^\ell} - \sum_{i=0}^{\ell-1} \alpha^i (1 + \alpha^{\ell-i} x) (1+x)^{\alpha^\ell - \alpha^{\ell-i}} \quad (2)$$

On en déduit que :

* $\text{SOC}(\ell, 0) = \text{SOC}(\ell, 1) = 0 \quad \forall \ell$;

* $\text{SOC}(1, n) = 0 \quad \forall n$;

* $\text{SOC}(\ell, n) = \frac{\alpha^{\ell-1}}{\alpha-1} \binom{\alpha^\ell}{n} - \sum_{i=1}^{\ell-1} \alpha^i \binom{\alpha^\ell - \alpha^{\ell-i}}{n} + \alpha^\ell \binom{\alpha^\ell - \alpha^{\ell-1}}{n-1}$

pour $n \geq 2$ et $\ell \geq 2$

(Le terme en $i=0$ est nul dès que $n \geq 2$).

Désormais, nous ne nous intéresserons qu'au cas $n, \ell \geq 2$;

On a alors

$$\overline{\text{SOC}}(\ell, n) = \frac{\alpha^{\ell-1}}{\alpha-1} - \sum_{i=1}^{\ell-1} \alpha^i \left[\frac{\binom{\alpha^{\ell}-\alpha^{\ell-i}}{n}}{\binom{\alpha^{\ell}}{n}} + \alpha^{\ell} \frac{\binom{\alpha^{\ell}-\alpha^{\ell-i}}{n-1}}{\binom{\alpha^{\ell}}{n}} \right] \quad (3)$$

Nous allons en proposer une valeur approchée, plus "parlante", en supposant que $n \ll \alpha^{\ell}$ (lire n très inférieur à α^{ℓ})

$$\text{Alors } \forall i \in \{1, \dots, \ell\} \quad \binom{\alpha^{\ell}-\alpha^{\ell-i}}{n} \simeq \frac{(\alpha^{\ell} - \alpha^{\ell-i})n}{n!}$$

D'où :

$$\begin{aligned} \overline{\text{SOC}}(\ell, n) &\simeq \frac{\alpha^{\ell-1}}{\alpha-1} - \sum_{i=1}^{\ell-1} \alpha^i \left(1 - \frac{1}{\alpha^i}\right)^n - \sum_{i=1}^{\ell-1} n \left(1 - \frac{1}{\alpha^i}\right)^{n-1} \\ &\simeq \frac{\alpha^{\ell-1}}{\alpha-1} - \sum_{i=1}^{\ell-1} \alpha^i \sum_{k=0}^n (-1)^k \binom{n}{k} \frac{1}{\alpha^{ik}} \\ &\quad - \sum_{i=1}^{\ell-1} n \sum_{k=0}^{n-1} (-1)^k \binom{n-1}{k} \frac{1}{\alpha^{ik}} \\ &\simeq \frac{\alpha^{\ell-1}}{\alpha-1} - \sum_{k=0}^n (-1)^k \binom{n}{k} \sum_{i=1}^{\ell-1} \frac{1}{\alpha^{i(k-1)}} \\ &\quad - \sum_{k=0}^{n-1} (-1)^k \binom{n-1}{k} n \sum_{i=1}^{\ell-1} \frac{1}{\alpha^{ik}} \\ &\simeq \frac{\alpha^{\ell-1}}{\alpha-1} \\ &\quad - \left(\left(\frac{\alpha^{\ell-1}}{\alpha-1} - 1 \right) - n(\ell-1) + \sum_{k=2}^n (-1)^k \binom{n}{k} \left[\frac{1 - \frac{1}{\alpha^{\ell(k-1)}}}{1 - \frac{1}{\alpha^{k-1}}} - 1 \right] \right) \\ &\quad - \left(n(\ell-1) + \sum_{k=1}^{n-1} n(-1)^k \binom{n-1}{k} \left[\frac{1 - \frac{1}{\alpha^{\ell k}}}{1 - \frac{1}{\alpha^k}} - 1 \right] \right) \end{aligned}$$

D'où pour n et ℓ plus grands que 2, pour $n \ll \alpha^\ell$ on a :

$$\overline{\text{SOC}}(\ell, n) \simeq 1 + \sum_{k=2}^n (-1)^k \binom{n}{k} (k-1) \left[\frac{1 - \frac{1}{\alpha^{\ell(k-1)}}}{1 - \frac{1}{\alpha^{k-1}}} - 1 \right] \quad (4)$$

Le nombre moyen de noeuds internes de l'ALC représentant n clés de ℓ caractères (pour n et ℓ plus grands que 2) vaut donc 1 (correspondant au sommet de l'arbre), augmenté d'une série alternée dont on donnera une valeur approchée).

La formule (4), est donc une approximation de la formule (3), un peu plus facile à interpréter.

Nous allons néanmoins en obtenir une autre, plus facile encore à l'interprétation en faisant l'hypothèse que les clés sont de longueur infinie.

Pour cela nous allons définir de nouvelles notations, puis revenir à la formule (1).

Soit :

$$\begin{aligned} * \quad \overline{\text{SOC}}(\infty, n) &= \lim_{\ell \rightarrow \infty} \overline{\text{SOC}}(\ell, n) \quad (\text{si cette limite existe}) \\ * \quad \text{SOC}^\infty(x) &= \sum_{n=0}^{\infty} \overline{\text{SOC}}(\infty, n) \cdot \frac{x^n}{n!} ; \end{aligned}$$

Alors [Flajolet, Sotteau], on a la propriété :

$$\text{SOC}^\infty(x) = \lim_{\ell \rightarrow \infty} \text{SOC}^{(\ell)}(x/\alpha^\ell)$$

De (1) on tire que :

$$\text{SOC}^{(\ell)}(x/\alpha^\ell) = \mathcal{P}^{(\ell)}(x/\alpha^\ell) - 1 - x + \alpha \mathcal{P}^{(\ell-1)}(x/\alpha^\ell)^{\alpha-1} \text{SOC}^{(\ell-1)}(x/\alpha^\ell)$$

En faisant tendre ℓ vers $+\infty$, il vient :

$$\text{SOC}^{(\infty)}(x) = \mathcal{P}^\infty(x) - 1 - x + \alpha \mathcal{P}^{(\infty)}(x/\alpha)^{\alpha-1} \text{SOC}^\infty(x/\alpha)$$

Mais $\mathcal{P}^{(\infty)}(x)$ est la série génératrice associée au paramètre valant 1 partout. En effet :

$$\mathcal{P}^{(\ell)}(x) = \sum_{p \in \mathcal{P}^{(\ell)}} 1 * x^{|p|} = \sum_{n=0}^{\alpha^{\ell}} \binom{\alpha^{\ell}}{n} x^n$$

$$\Rightarrow p(\ell, n) = 1 \quad \forall n, \ell ;$$

D'où $p(\infty, n) = 1 \quad \forall n \Rightarrow \mathcal{P}^{\infty}(x) = \sum_{n \geq 0} 1 * \frac{x^n}{n!} = e^x$

On en déduit que :

$$\text{SOC}^{(\infty)}(x) = e^x - 1 - x + \alpha e^{x(1-\alpha^{-1})} * \text{SOC}^{\infty}(x/\alpha)$$

Par récurrence, il vient :

$$\begin{aligned} \text{SOC}^{(\infty)}(x) &= \sum_{k=0}^{\infty} (e^{x/\alpha^k} - 1 - x/\alpha^k) \left(\prod_{i=1}^{k-1} \alpha e^{(x/\alpha^i)(1-\alpha^{-1})} \right) \\ &= \sum_{k=0}^{\infty} \alpha^k (e^{x/\alpha^k} - 1 - x/\alpha^k) e^{x(1-\alpha^{-k})} \\ \text{SOC}^{(\infty)}(x) &= \sum_{k=0}^{\infty} \alpha^k (e^x - e^{x(1-\alpha^{-k})} - x/\alpha^k e^{x(1-\alpha^{-k})}) ; \end{aligned} \quad (5)$$

$$\text{D'où } \overline{\text{SOC}}^{(\infty, n)} = \sum_{k=0}^{\infty} \alpha^k (1 - (1-\alpha^{-k})^n - n\alpha^{-k}(1-\alpha^{-k})^{n-1}) \quad (6)$$

$$\text{Soit } \varphi(x) = \sum_{k \geq 0} \alpha^k (1 - e^{-x\alpha^{-k}} - x\alpha^{-k} e^{-x\alpha^{-k}})$$

On montre que $\overline{\text{SOC}}^{(\infty, n)} - \varphi(n)$ est majorée par une fonction linéaire en $O(n^{2/3})$:

$$\overline{\text{SOC}}^{(\infty, n)} - \varphi(n) = O(n^{2/3}) \quad (\text{voir annexe 1}) \quad (8)$$

Notons $\varphi^*(s) = \int_0^{\infty} \varphi(x) x^{s-1} dx$ la transformée de Mellin de φ en s , si cette intégrale à un sens.

On vérifie que $\Psi^*(s) = - \frac{(s+1)\Gamma(s)}{1-\alpha^{s+1}}$ (voir Annexe 2) (9)

où $\Gamma(s) = \int_0^\infty e^{-x} x^{s-1} dx$

En utilisant le théorème d'inversion de la transformation de Mellin, il vient :

$$\begin{aligned} \Psi(x) &= \frac{1}{2i\pi} \int_{c-i\infty}^{c+i\infty} \Psi^*(s) x^{-s} ds \\ &= \frac{x}{\log \alpha} - x \sum_{h \in \mathbb{Z} - \{0\}} \frac{2ih\pi}{(\log \alpha)^2} \Gamma(-1+2ih\pi) e^{-2ikh\pi \log_\alpha(x)} \\ &\quad + O(|x|^{1/2}) \quad (\text{voir Annexe 3}) \end{aligned} \quad (10)$$

D'où en cumulant les résultats :

$$\begin{aligned} \overline{\text{SOC}}(\infty, n) &= \frac{n}{\log \alpha} - n \left[\sum_{k \in \mathbb{Z} - \{0\}} \frac{2ik\pi}{(\log \alpha)^2} P(-1+2ik\pi) e^{-2ik\pi \log_\alpha(x)} \right] \\ &\quad + O(n^{2/3}) + O(n^{1/2}) \end{aligned}$$

On montre que pour des valeurs raisonnables des paramètres α , la série de Fourier est numériquement négligeable. (Un ordre de grandeur est 10^{-2} au plus). Il vient :

$$\overline{\text{SOC}}(\infty, n) = \frac{n}{\log \alpha} + O(n^{2/3})$$

CALCUL DU NOMBRE MOYEN DE SOMMETS N'AYANT QU'UN FILS, DANS UN ARC :

En reprenant des notations identiques à celles utilisées pour $\text{SOC}(\ell, B)$, on a : Pour $p \in \mathcal{G}^{(\ell)}$:

$$\begin{aligned} \text{SOC1}(p) &= \sum_{i=1}^{\alpha} \text{SOC1}(p_i) \\ &\quad + \sum_{i=1}^{\alpha} \mathbb{1}_{\{|p_i| \geq 2\}} \left(\prod_{\substack{k=1 \\ k \neq i}}^{\alpha} \mathbb{1}_{\{p_k \neq \emptyset\}} \right) \end{aligned} \quad (o)$$

(En effet le sommet de l'arbre est un noeud interne n'ayant qu'un fils si l'arbre a un sous arbre contenant plus de deux chaînes de caractères et tous ses autres sous arbres vides).

On en déduit

$$\text{SOCl}^{(\ell)}(x) = \alpha \left(\mathcal{P}^{(\ell-1)}(x) \right)^{\alpha-1} \text{SOCl}^{(\ell-1)}(x) + \alpha \mathcal{P}^{(\ell-1)}(x) - 1 - \alpha^{\ell-1} x \quad (1)$$

Par récurrence, il vient

$$\text{SOCl}^{(\ell)}(x) = \sum_{i=0}^{\ell-1} \alpha^{i+1} \left[(1+x)^{\alpha^{\ell} + \alpha^{\ell-i-1} - \alpha^{\ell-i}} - (1+\alpha^{\ell-i-1}x)(1+x)^{\alpha^{\ell} - \alpha^{\ell-i}} \right] \quad (2)$$

En extrayant le coefficient d'ordre n et en divisant par $\binom{\alpha^{\ell}}{n}$, il vient :

$$\begin{aligned} \overline{\text{SOCl}}(\ell, n) = \sum_{i=0}^{\ell-1} \alpha^{i+1} \frac{\binom{\alpha^{\ell} + \alpha^{\ell-i-1} - \alpha^{\ell-i}}{n}}{\binom{\alpha^{\ell}}{n}} - \sum_{i=1}^{\ell-1} \alpha^{i+1} \frac{\binom{\alpha^{\ell} - \alpha^{\ell-i}}{n}}{\binom{\alpha^{\ell}}{n}} \\ - \sum_{i=1}^{\ell-1} \alpha^{\ell} \frac{\binom{\alpha^{\ell} - \alpha^{\ell-i}}{n-1}}{\binom{\alpha^{\ell}}{n}} \end{aligned} \quad (3)$$

En supposant $n \ll \alpha^{\ell}$, (3) devient :

$$\overline{\text{SOCl}}(\ell, n) \simeq \sum_{k=2}^n (-1)^k \binom{n}{k} \alpha \left[(1-\alpha^{-1})^{k-1+k\alpha} - 1 \right] \left[\frac{1 - \frac{1}{\alpha^{\ell(k-1)}}}{1 - \frac{1}{\alpha^{k-1}}} - 1 \right] \quad (4)$$

Si l'on fait tendre ℓ vers l'infini dans (1), on obtient :

$$\text{SOCl}^{(\infty)}(x) = \sum_{k \geq 0} \alpha^{k+1} \left[\exp \left[x(1-\alpha^{-k} + \alpha^{-k-1}) \right] - \exp \left[x(1-\alpha^{-k}) \right] - x\alpha^{-k-1} \exp \left[x(1-\alpha^{-k}) \right] \right] \quad (5)$$

D'où l'on extraie le coefficient d'ordre n :

$$\overline{\text{SOCl}}(\infty, n) = \sum_{k \geq 0} \alpha^{k+1} ((1-\alpha^{-k} + \alpha^{-k-1})^n - (1-\alpha^{-k})^n - n\alpha^{-k-1}(1-\alpha^{-k})^{n-1}) \quad (6)$$

$$\text{Soit } \varphi(x) = \sum_{k \geq 0} \alpha^{k+1} (\exp[x(\alpha^{-k} + \alpha^{-k-1})] - \exp[x\alpha^{-k}] - x\alpha^{-k-1} \exp[x\alpha^{-k}]) \quad (7)$$

On montre que :

$$\overline{\text{SOC1}}(\infty, n) - \varphi(n) = O(n^{2/3}) \quad (8)$$

puis que

$$\varphi^*(s) = \frac{\alpha \Gamma(s) ((1-\alpha^{-1})^{-s-1} - s\alpha^{-1})}{1 - \alpha^{s+1}} \quad (9)$$

En utilisant le théorème d'inversion, on en déduit :

$$\begin{aligned} \varphi(x) &= \frac{x}{\log \alpha} \left[\log(1 - \frac{1}{\alpha}) * (\alpha-1) + 1 \right] \\ &\quad - x \sum_{k \in \mathbb{Z} - \{0\}} c_k e^{-2ik\pi \log_{\alpha}(x)} \\ &\quad + O(|x|^{1/2}) \end{aligned} \quad (10)$$

$$\text{où } c_k = \frac{\alpha}{\log \alpha} \Gamma(-1 + \frac{2ik\pi}{\log \alpha}) \left[(1-\alpha^{-1})^{1 - \frac{2ik\pi}{\log \alpha}} - 1 + \alpha^{-1} (1 - \frac{2ik\pi}{\log \alpha}) \right]$$

On obtient donc finalement :

$$\overline{\text{SOC1}}(\infty, n) = \frac{x}{\log \alpha} \left[1 + (\alpha-1) \log(1-\alpha^{-1}) \right] + O(n^{2/3})$$

CALCUL DU NOMBRE DE SOMMETS D'UN AL

Pour $p \in \mathfrak{P}^{(\ell)}$, on a

$$\text{SOM}(p) = \mathbb{1}_{\{p \neq \emptyset\}} + \sum_{i=1}^{\alpha} \text{SOM}(p_i) \quad (o)$$

$$\text{D'où } \text{SOM}^{(\ell)}(x) = \mathfrak{P}^{(\ell)}(x) - 1 + \alpha (\mathfrak{P}^{(\ell-1)}(x))^{\alpha-1} \text{SOM}^{(\ell-1)}(n) \quad (1)$$

D'où l'on extraie le coefficient d'ordre n , et en divisant par $\binom{\alpha \ell}{n}$ on obtient :

$$\overline{\text{SOM}}(\ell, n) = \alpha^\ell \frac{\binom{\alpha^\ell - 1}{n-1}}{\binom{\alpha^\ell}{n}} + \frac{\alpha^{\ell-1}}{\alpha-1} \binom{\alpha^\ell}{n} + \sum_{i=1}^{\ell-1} \alpha^i \frac{\binom{\alpha^\ell - \alpha^{\ell-i}}{n}}{\binom{\alpha^\ell}{n}} \quad (3)$$

En faisant l'approximation $n \ll \alpha^\ell$, on en déduit :

$$\overline{\text{SOM}}(\ell, n) = 1 + n\ell - \sum_{k=2}^n (-1)^k \binom{n}{k} \left[\frac{1 - \frac{1}{\alpha^{\ell(k-1)}}}{1 - \frac{1}{\alpha^{k-1}}} - 1 \right] \quad (4)$$

$\text{SOM}(\ell, n)$ est donc essentiellement linéaire en $n\ell$ diminué d'une série $S(\ell, n)$ dont nous allons calculer une valeur approchée.

$$\begin{aligned} S(\ell, n) &= \sum_{k=2}^n (-1)^k \binom{n}{k} \sum_{i=1}^{\ell-1} \alpha^{-i(k-1)} \\ &= \sum_{i=1}^{\ell-1} \alpha^i \left[(1 - \alpha^{-i})^n - 1 + n\alpha^{-i} \right] \end{aligned}$$

Notons $X_i = \alpha^i \left[(1 - \alpha^{-i})^n - 1 + n\alpha^{-i} \right]$

$$S(\ell, n) = \left(\sum_{i \geq 0} X_i \right) - X_0 - \left(\sum_{i=\ell}^{\infty} X_i \right)$$

Or $X_0 = n - 1$;

$$\begin{aligned} \text{Par ailleurs } X_i &= \alpha^i \left[\exp(n(-\alpha^{-i} + \alpha^{-2i} O(1))) - 1 + n\alpha^{-i} \right] \\ &= \alpha^i \left[1 - n\alpha^{-i} + n\alpha^{-2i} O(1) - 1 + n\alpha^{-i} \right] \\ &= n\alpha^{-i} O(1) \end{aligned}$$

Donc $\sum_{i=\ell}^{\infty} X_i = \sum_{i=\ell}^{\infty} n\alpha^{-i} O(1) = n\alpha^{-\ell} O(1) = O(n\alpha^{-\ell})$

Donc $S(\ell, n) = \sum_{i=0}^{\infty} X_i - n + 1 + O(n\alpha^{-\ell})$

Notons

$$T(n) = \sum_{i=0}^{\infty} X_i$$

et

$$\varphi(x) = \sum_{i=0}^{\infty} \alpha^i \left[e^{-x\alpha^{-i}} - 1 + x\alpha^{-i} \right]$$

On montre que :

$$T(n) - \varphi(n) = O(n^{1/2})$$

(La démonstration est identique à celle de l'annexe 1, sinon que l'on prend $\varepsilon = 1/2$).

On calcule φ^* et l'on trouve :

$$\varphi^*(s) = \frac{\Gamma(s)}{1 - \alpha^{s+1}}$$

En utilisant le théorème d'inversion, il vient :

$$\begin{aligned} \varphi(x) &= x \frac{(\log x + \gamma - 1)}{\log \alpha} + x \sum \frac{\Gamma(-1 + \frac{2ik\pi}{\log \alpha}) e^{-2ik\pi \log_{\alpha} x}}{\log \alpha} \\ &\quad + O(|x|^{1/2}) \end{aligned}$$

D'où en négligeant la série de Fourier, on a donc :

$$\begin{aligned} S(\ell, n) &= n \frac{(\log n + \gamma - 1)}{\log \alpha} - n + 1 + O(n\alpha^{-\ell}) + O(n^{1/2}) \\ &= n \left[\frac{\log n + \gamma - 1 - \log \alpha}{\log \alpha} \right] + O(n^{1/2}) \end{aligned}$$

(Comme $n < \alpha^{\ell}$, on a $n\alpha^{-\ell} < 1$; γ représente ici la constante d'Euler)

CALCUL DU NOMBRE MOYEN DE SOMMETS N'AYANT QU'UN FILS DANS UN AL :

Soit $p \in \mathcal{P}^{(\ell)}$;

$$SOM1(p) = \sum_{i=1}^{\alpha} \mathbb{1}_{\{p_i \neq 0\}} \prod_{\substack{j=1 \\ j \neq i}}^{\alpha} \mathbb{1}_{\{p_j \neq 0\}} + \sum_{i=1}^{\alpha} SOM1(p_i) \quad (o)$$

$$D'o\grave{u} \text{ SOMI}^{(\ell)}(x) = \alpha (\mathcal{P}^{(\ell-1)}(x) - 1) + \alpha (\mathcal{P}^{(\ell-1)}(x))^{\alpha-1} \text{SOMI}^{(\ell-1)}(x) \quad (1)$$

Par r  currence on en d  duit que :

$$\text{SOMI}^{(\ell)}(x) = \sum_{i=0}^{\ell-1} \alpha^{i+1} \left[(1+x)^{\alpha^{\ell}-\alpha^{\ell-i}+\alpha^{\ell-i-1}} - (1+x)^{\alpha^{\ell}-\alpha^{\ell-i}} \right] \quad (2)$$

D'o   l'on d  duit la valeur exacte de $\overline{\text{SOMI}}(\ell, n)$:

$$\overline{\text{SOMI}}(\ell, n) = \sum_{i=0}^{\ell-1} \alpha^{i+1} \left[\frac{\binom{\alpha^{\ell}-\alpha^{\ell-i}+\alpha^{\ell-i-1}}{n}}{\binom{\alpha^{\ell}}{n}} - \frac{\binom{\alpha^{\ell}-\alpha^{\ell-i}}{n}}{\binom{\alpha^{\ell}}{n}} \right] \quad (3)$$

(pour n et ℓ plus grand que 2, et avec la convention :

$$\binom{n}{k} = 0 \text{ si } k > n).$$

Si l'on suppose, $n \ll \alpha^{\ell}$, on d  duit de (3) que :

$$\overline{\text{SOMI}}(\ell, n) = \sum_{i=0}^{\ell-1} \alpha^{i+1} \left[(1-\alpha^{-i}(1-\alpha^{-1}))^n - (1-\alpha^{-i})^n \right]$$

qui s'  crit encore

$$\text{SOMI}(\ell, n) = n\ell - \sum_{k=2}^n (-1)^k \binom{n}{k} (1-(1-\alpha^{-1})^k) \left(\frac{1-\alpha^{-\ell(k-1)}}{1-\alpha^{-(k-1)}} \right) \quad (4)$$

Comme $\text{SOM}(\ell, n)$, $\text{SOMI}(\ell, n)$ est une fonction essentiellement lin  aire en $n\ell$, diminu  e d'une s  rie $S(\ell, n)$ dont nous allons calculer une valeur approch  e ,

$$\begin{aligned} S(\ell, n) &= \sum_{i=0}^{\ell-1} \alpha^{i+1} \sum_{k=2}^n (-1)^k \binom{n}{k} \alpha^{-ik} \left[1 - (1-\alpha^{-1})^k \right] \\ &= \sum_{i=0}^{\ell-1} \alpha^{i+1} \left[(1-\alpha^{-i})^n - 1 + n\alpha^{-i} - (1-\alpha^{-i}(1-\alpha^{-1}))^n + 1 - n\alpha^{-i}(1-\alpha^{-1}) \right] \\ &= \sum_{i=0}^{\ell-1} \alpha^{i+1} \left[(1-\alpha^{-i})^n - (1 - \alpha^{-i}(1-\alpha^{-1}))^n + n\alpha^{-i-1} \right] \\ &= \sum_{i=0}^{\ell-1} X_i \end{aligned}$$

De même que pour $\overline{\text{SOM}}(\ell, n)$, on montre que

$$S(\ell, n) = \sum_{i=0}^{\infty} X_i + O(n\alpha^{-\ell})$$

Notons

$$\varphi(x) = \sum_{i=0}^{\infty} \alpha^{i+1} \left[e^{-x\alpha^{-i}} - e^{-x\alpha^{-i}(1-\alpha^{-1})} + x\alpha^{-i-1} \right]$$

On montre que

$$\sum_{i=0}^{\infty} X_i - \mathcal{C}(n) = O(n^{1/2})$$

On calcule alors φ^* et l'on trouve

$$\varphi^*(s) = \frac{\alpha \Gamma(s) [1 - (1-\alpha^{-1})^{-s}]}{1-\alpha^{s+1}}$$

En appliquant le théorème d'inversion, on trouve :

$$\varphi(s) = x \left[\frac{\log x - (\alpha - 1) \log(1-\alpha^{-1}) + (\gamma-1)}{\log \alpha} \right]$$

$$+ x \sum_{\substack{k \in \mathbb{Z} \\ k \neq 0}} X_k e^{-2ik\pi \log_{\alpha} x} + O(|x|^{1/2})$$

où

$$C_k = \frac{\alpha \Gamma(-1 + \frac{2ik\pi}{\log \alpha}) \left[1 - (1-\alpha^{-1})^{1 - \frac{2ik\pi}{\log \alpha}} \right]}{\log \alpha}$$

D'où l'on tire que $\overline{\text{SOM}}(\ell, n)$ vaut essentiellement

$$\begin{aligned} \overline{\text{SOM}}(\ell, n) &= n\ell - n \frac{\log n - (\alpha-1)\log(1-\alpha^{-1}) + (\gamma-1)}{\log \alpha} \\ &\quad + O(n^{1/2}) \end{aligned}$$

(où γ désigne la constante d'Euler).

Pour récapituler les résultats, nous avons montré que, pour $n, \ell \geq 2$:

1) Sous hypothèse de longueur fixe ℓ :

$$\overline{\text{SOC}}(\ell, n) = \frac{\alpha^{\ell-1}}{\alpha-1} - \sum_{i=1}^{\ell-1} \left[\alpha^i \frac{\binom{\alpha^{\ell}-\alpha^{\ell-i}}{n}}{\binom{\alpha^{\ell}}{n}} - \alpha^{\ell} \frac{\binom{\alpha^{\ell}-\alpha^{\ell-i}}{n}}{\binom{\alpha^{\ell}}{n}} \right]$$

$$\overline{\text{SOC1}}(\ell, n) = \sum_{i=0}^{\ell-i} \left[\alpha^{i+1} \frac{\binom{\alpha^{\ell}-\alpha^{\ell-i}+\alpha^{\ell-i-1}}{n}}{\binom{\alpha^{\ell}}{n}} - \alpha^{i+1} \frac{\binom{\alpha^{\ell}-\alpha^{\ell-i}}{n}}{\binom{\alpha^{\ell}}{n}} - \alpha^{\ell} \frac{\binom{\alpha^{\ell}-\alpha^{\ell-i}}{n-1}}{\binom{\alpha^{\ell}}{n}} \right]$$

$$\overline{\text{SOM}}(\ell, n) = n + \frac{\alpha^{\ell-1}}{\alpha-1} + \sum_{i=1}^{\ell-1} \alpha^i \frac{\binom{\alpha^{\ell}-\alpha^{\ell-i}}{n}}{\binom{\alpha^{\ell}}{n}}$$

$$\overline{\text{SOM1}}(\ell, n) = \sum_{i=0}^{\ell-1} \alpha^{i+1} \left[\frac{\binom{\alpha^{\ell}-\alpha^{\ell-i}+\alpha^{\ell-i-1}}{n}}{\binom{\alpha^{\ell}}{n}} - \frac{\binom{\alpha^{\ell}-\alpha^{\ell-i}}{n}}{\binom{\alpha^{\ell}}{n}} \right]$$

2) Sous hypothèse de longueur fixe, en supposant $\alpha^{\ell} \gg n$:

$$\overline{\text{SOC}}(\ell, n) = 1 + \sum_{k=2}^n (-1)^k \binom{n}{k} (k-1) \left[\frac{1 - \alpha^{-\ell(k-1)}}{1 - \alpha^{-(k-1)}} - 1 \right]$$

$$\overline{\text{SOC1}}(\ell, n) = \sum_{k=2}^n (-1)^k \binom{n}{k} \left[(1-\alpha^{-1})^k - 1 + k\alpha^{-1} \right] \left[\frac{1 - \alpha^{-\ell(k-1)}}{1 - \alpha^{-(k-1)}} \right]$$

$$\overline{\text{SOM}}(\ell, n) = 1 + n\ell - \sum_{k=2}^n (-1)^k \binom{n}{k} \left[\frac{1 - \alpha^{-\ell(k-1)}}{1 - \alpha^{-(k-1)}} - 1 \right]$$

$$\overline{\text{SOM1}}(\ell, n) = n\ell - \sum_{k=2}^n (-1)^k \binom{n}{k} \left[1 - (1-\alpha^{-1})^k \right] \left[\frac{1 - \alpha^{-\ell(k-1)}}{1 - \alpha^{-(k-1)}} \right]$$

3) En faisant l'approximation $\ell = \infty$

$$\overline{\text{SOC}}(\ell, n) = \frac{n}{\log \alpha} + O(n^{2/3})$$

$$\overline{\text{SOC1}}(\ell, n) = \frac{n}{\log \alpha} \left[1 + (\alpha-1) \log(1-\alpha^{-1}) \right] + O(n^{2/3})$$

4) En approchant les formules de 2) :

$$\overline{SOM}(\ell, n) = 1 + n\ell - \frac{n}{\log \alpha} (\log n + \gamma - 1 - \log \alpha) + O(n^{1/2})$$

$$\overline{SOMl}(\ell, n) = n\ell - \frac{n}{\log \alpha} (\log n - (\bar{\alpha}-1) \log(1-\alpha^{-1}) + \gamma - 1) + O(n^{1/2})$$

4. L'APPLICATION VERSO

VERSO est une machine base de données relationnelle dorsale, développée à l'INRIA [VERSO]. Sa principale caractéristique est un filtre câblé de type automate à états finis qui traite les données au rythme de deux millions d'octets par seconde. Ce filtre effectue "au vol" pendant leur transfert de la mémoire secondaire vers la mémoire centrale, toutes les opérations unaires de l'algèbre relationnelle (insertion, suppression, sélection, projection), et les opérations binaires sur des relations physiquement ordonnées (union, intersection, différence, jointure, division).

L'espace secondaire de VERSO est un disque dur de 300 Mega-Octets, soit 15 K pistes de 20 K octets. Traditionnellement, l'unité adressable (ou le bloc) est le secteur de piste, dont la taille varie entre 512 et 4 K-octets. Comme nous disposons d'un processeur très rapide, il est préférable d'avoir des unités adressables beaucoup plus grosses (compromis temps de calcul, temps d'accès disque) : le bloc de VERSO est la piste entière soit 20 K octets.

L'index contient donc 15 K n-uplets au plus.

Nous utilisons maintenant les résultats des formules 3) et 4) de la section 3 pour représenter les variations de la taille de l'index en fonction des divers paramètres du modèle. Considérons la relation $R[A_1, A_2, A_3]$. La figure 9 représente la taille de l'index en fonction de celle de l'alphabet pour diverses valeurs de la sélectivité, et des attributs de longueur moyenne 10.

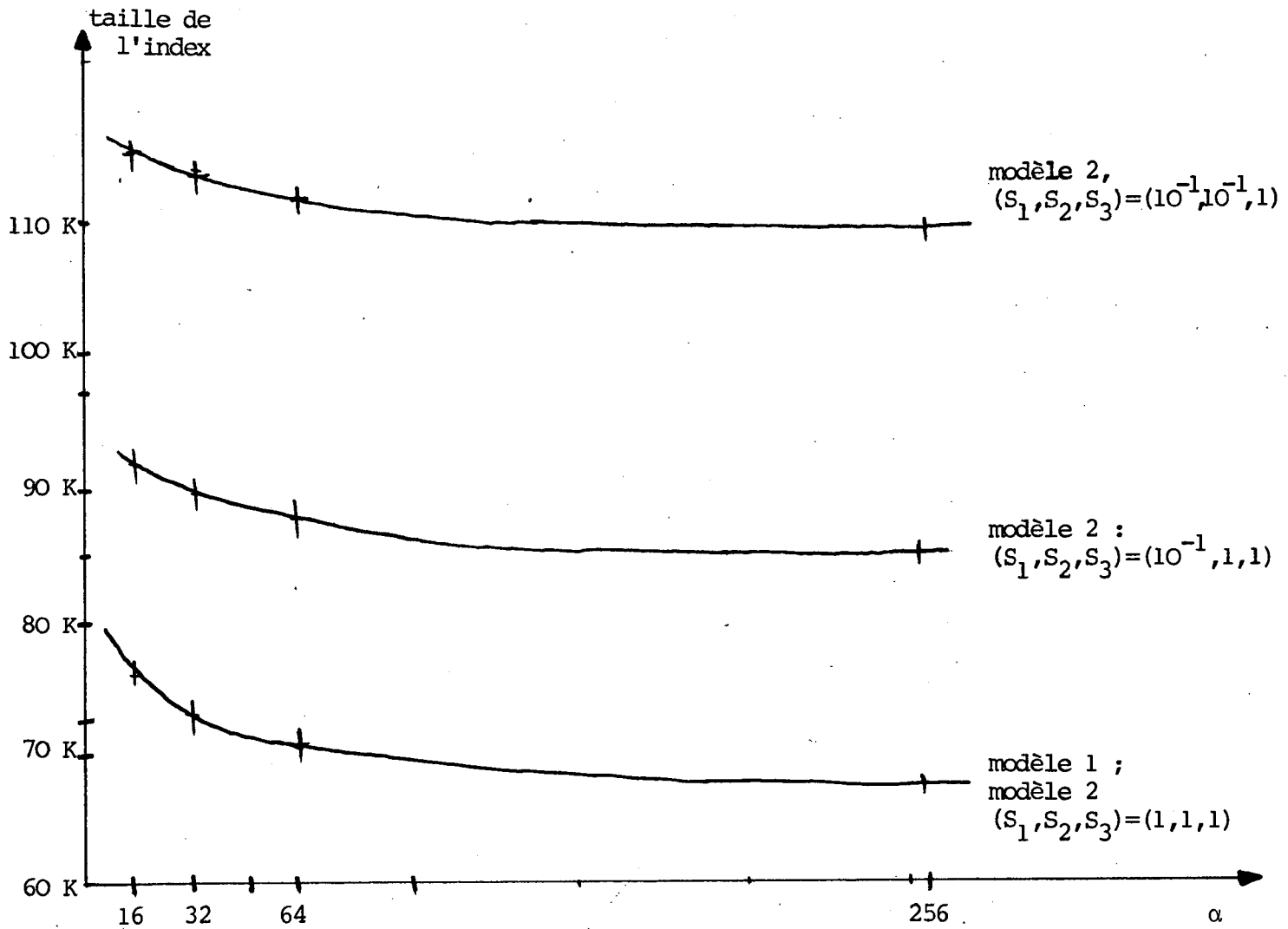


Figure 9 : Taille de l'index en fonction de la taille de l'alphabet

Les figures 10-a, 10-b, 10-c, représentent les variations de la taille de l'index en fonction de S_1 pour des valeurs respectives de S_2 1, 10^{-1} , 10^{-2} . La longueur des attributs restent à 10 caractères et l'alphabet est de taille 32.

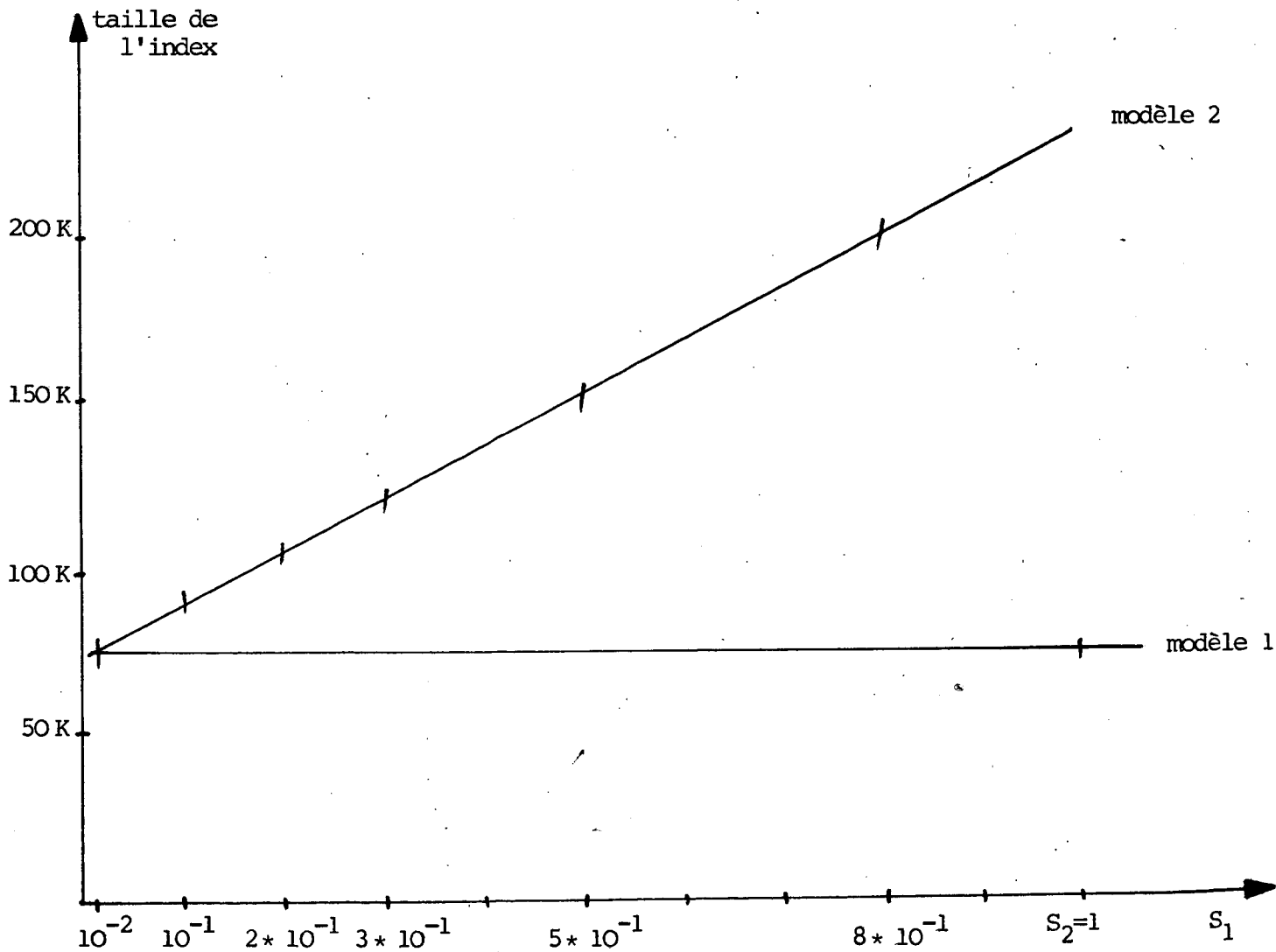


Figure 10-a : Taille de l'index en fonction de S_1 pour $S_2=1$

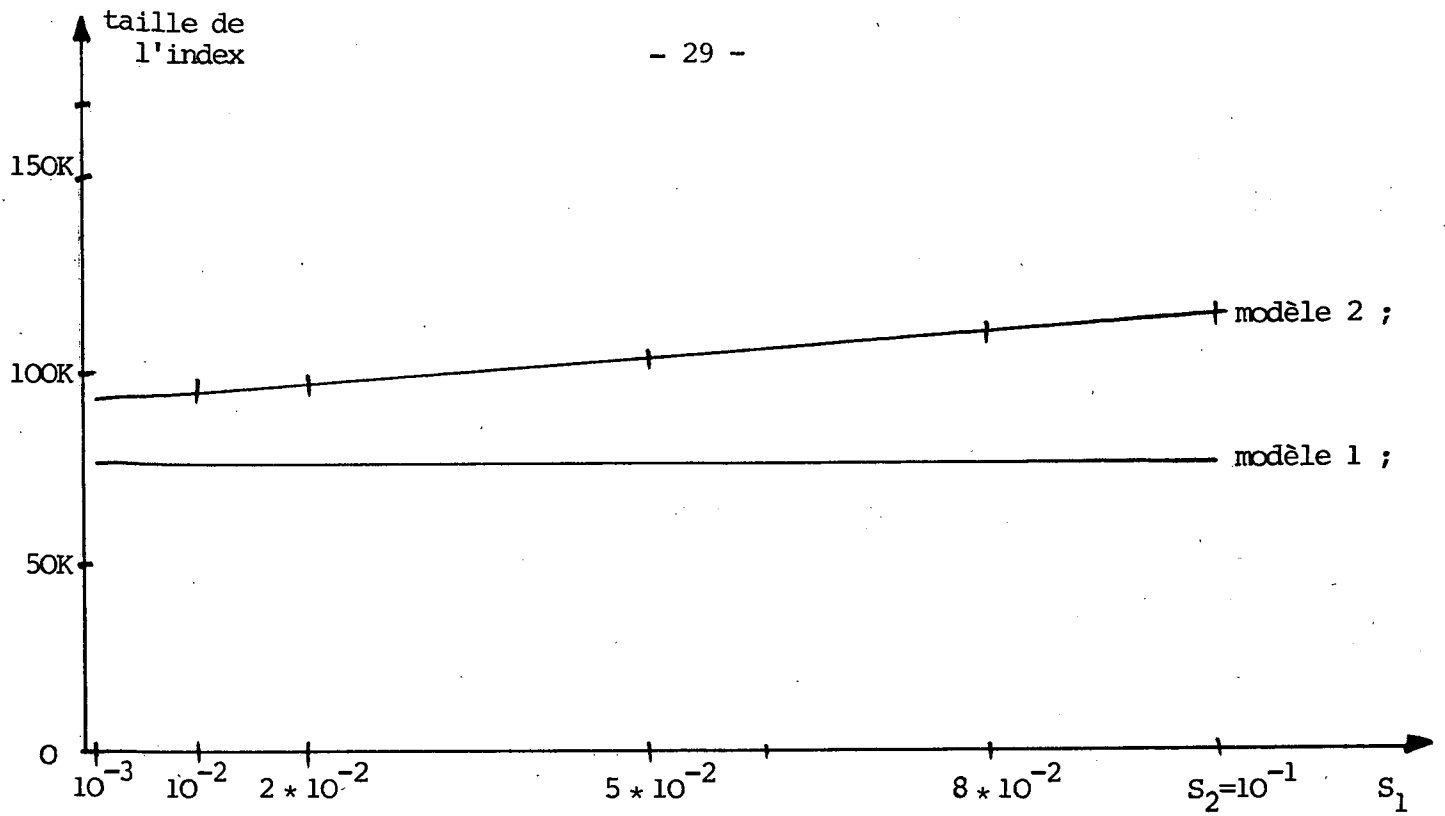


Figure 10-b : Taille de l'index en fonction de S_1 pour $S_2 = 10^{-1}$

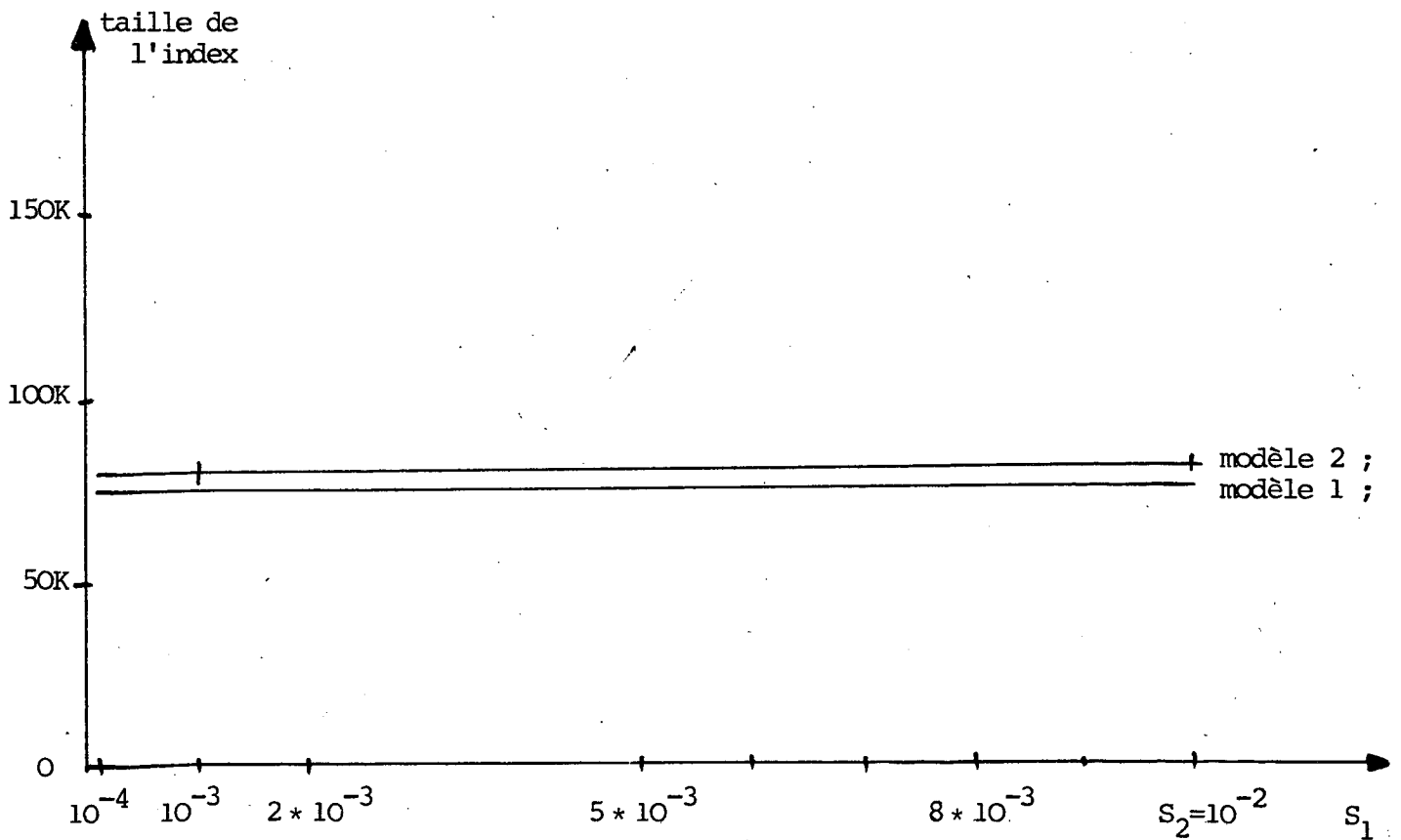


Figure 10-c : Taille de l'index en fonction de S_1 pour $S_2 = 10^{-2}$

Nous constatons qu'il y a une forte différence dans les résultats donnés par le modèle 1 et le modèle 2 pour les valeurs de S_1 voisine de 1. Le modèle 2 donne de trop grandes valeurs dans ce cas.

Pour finir, la figure 11 représente la taille de l'index en fonction de la longueur des attributs (nous supposons les attributs de même longueur) pour un alphabet de taille 32 et diverses sélectivités.

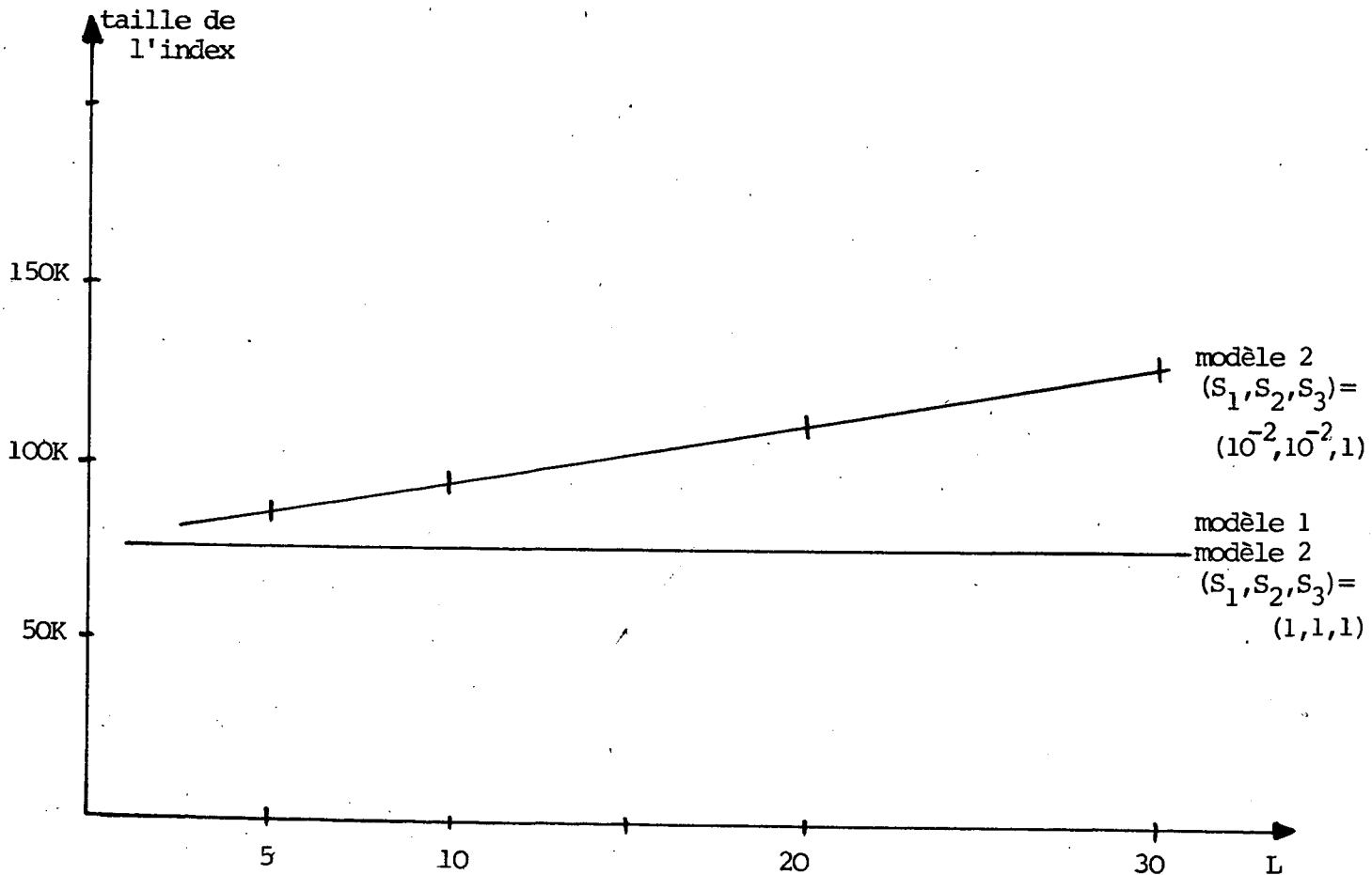


Figure 11 : Taille de l'index en fonction de la taille des attributs

Pour résumer, nous constatons que la taille de l'index croît faiblement en fonction de celle de l'alphabet. Elle croît linéairement en fonction de la taille des attributs sous le modèle 2 mais reste invariante sous le modèle 1.

Enfin, elle croît linéairement en fonction de la sélectivité sous le modèle 2, et reste constante sous le modèle 1. Notons que le modèle 2 donne des résultats numériquement supérieurs à ceux du modèle 1. Il est impossible d'établir lequel est le meilleur sinon en faisant des mesures. En conclusion, nous obtenons un index de l'ordre de 100 K octets pour un espace secondaire de 300 Mega-octets. Examinons un index de type arbre B de degré m : si l'on suppose que l'arbre est aussi compact que possible (i.e. chaque noeud de l'arbre a m fils); on en déduit que la hauteur minimum de l'arbre est

$$h = \frac{\log N}{\log m} \quad (N \text{ est le nombre de clés})$$

D'où le nombre de noeuds de l'arbre est au moins :

$$\text{nb noeuds} \geq \frac{m^h - 1}{m - 1} = \frac{N - 1}{m - 1}$$

La taille d'un noeud est $m * p + (m-1) * \ell$ où p est la taille d'un pointeur (ou encore l'adresse d'un bloc sur mémoire secondaire) et ℓ la longueur moyenne d'un n -uplet.

En prenant $p=2$ octets et $\ell=30$ octets, si l'on veut que chaque noeud occupe un bloc sur le disque, on a :

$$32 m \simeq 20 K \Rightarrow m \simeq 625$$

$$\text{D'où } \text{Nb noeuds} \geq \frac{15000 - 1}{625 - 1} \simeq 24$$

La taille de l'index est donc au moins de 480 K. Dans le cadre de VERSO, notre structure nous permet de maintenir l'index résident. Afin d'accélérer la manipulation de l'index, un deuxième niveau est nécessaire, dont la taille ne dépasse pas une centaine d'octets.

5. CONCLUSION

Nous avons décrit une structure de type arbre lexicographique, qui permet d'indexer un fichier totalement ordonné. La propriété de cette structure est d'être suffisamment compacte pour maintenir l'index résident en mémoire centrale. Pour un espace secondaire de 300 Mégaoctets divisés en 15 K unités adressables, nous avons évalué la taille de l'index à 100 K octets environ (c'est un ordre de grandeur plus qu'une valeur précise), alors qu'une structure de type arbre-B nous amène à une taille de l'ordre de 500 K octets.

ANNEXE 1 Preuve de $\overline{\text{SOC}}(\infty, n) - \varphi(n) = O(n^{2/3})$

$$\overline{\text{SOC}}(\infty, n) - \varphi(n) = \sum_{k \geq 0} x_k + y_k$$

où $x_k = \alpha^k (e^{-n\alpha^{-k}} - (1 - \alpha^{-k})^n)$

et $y_k = n (e^{-n\alpha^{-k}} - (1 - \alpha^{-k})^{n-1})$

Posons $x = n\alpha^{-k}$; alors :

$$x_k = \frac{n}{x} (e^{-x} - (1 - \frac{x}{n})^n)$$

$$y_k = n (e^{-x} - (1 - \frac{x}{n})^{n-1})$$

Soit $\epsilon \in]0, 1[$;

1er cas : $x \leq n^\epsilon$ ($\Leftrightarrow k \geq (1-\epsilon) \log_\alpha n$)

Alors $(1 - \frac{x}{n})^n = \exp[n \log(1 - \frac{x}{n})] =$
 $= \exp[-x - O(x^2 n^{-1})]$

D'où $x_k = \frac{n}{x} e^{-x} [1 - \exp(-O(x^2 n^{-1}))]$
 $= \frac{n}{x} e^{-x} O(x^2 n^{-1})$

En remplaçant x par sa valeur et en majorant e^{-x} par 1, il vient :

$$x_k = \alpha^{-k} O(n)$$

D'où $\sum_{k \geq (1-\epsilon) \log_\alpha n} x_k = O(n) \sum_{k \geq (1-\epsilon) \log_\alpha n} \alpha^{-k}$

On en tire :

$$\sum_{k \geq (1-\epsilon) \log_\alpha n} x_k = \frac{O(n)}{\alpha^{(1-\epsilon) \log_\alpha n}} \sum_{k \geq 0} \alpha^{-k}$$

(Nous avons identifié $(1-\epsilon) \log_\alpha n$ avec sa partie entière, ce qui n'a pas d'importance car on s'intéresse à l'ordre de la série et non à sa valeur exacte. Cette approximation est répétée par la suite).

Finalement, nous obtenons :

$$\begin{aligned} \sum_{k \geq (1-\epsilon) \log_{\alpha} n} x_k &= \frac{O(n)}{n^{1-\epsilon}} \left(\sum_{k \geq 0} \alpha^{-k} \right) \\ &= O(n^{\epsilon}) \end{aligned}$$

De même :

$$\begin{aligned} y_k &= n[e^{-x} - e^{(n-1)[-xn^{-1} - O(x^2 n^{-2})]}] \\ &= ne^{-x} [1 - e^{xn^{-1} - O(n^2 n^{-1})}] \\ &= ne^{-x} [-xn^{-1} + O(x^2 n^{-1})] \\ &= e^{-n\alpha^{-k}} [n^2 \alpha^{-2k} - n\alpha^{-k}] O(1) \end{aligned}$$

D'où

$$\begin{aligned} \sum_{k \geq (1-\epsilon) \log_{\alpha} n} y_k &= n^2 O(1) \sum_{k \geq (1-\epsilon) \log_{\alpha} n} \alpha^{-2k} + n O(1) \sum_{k \geq (1-\epsilon) \log_{\alpha} n} \alpha^{-k} \\ &= \frac{n^2 O(1)}{\alpha^{2(1-\epsilon) \log_{\alpha} n}} \left(\sum_{k \geq 0} \alpha^{-2k} \right) + \frac{n O(1)}{\alpha^{(1-\epsilon) \log_{\alpha} n}} \left(\sum_{k \geq 0} \alpha^{-k} \right) \\ &= O(n^{2\epsilon}) + O(n^{\epsilon}) = O(n^{2\epsilon}) \end{aligned}$$

2ème cas : $x > n^{\epsilon}$; ($\Leftrightarrow k \leq (1-\epsilon) \log_{\alpha} n$)

$$\forall a > 0 \quad 0 < (1-a) < e^{-a}$$

$$\Rightarrow 0 < (1-\alpha^{-k}) < e^{-\alpha^{-k}}$$

$$\Rightarrow 0 < (1-\alpha^{-k})^n < e^{-n\alpha^{-k}}$$

$$\Rightarrow 0 < e^{-n\alpha^{-k}} - (1-\alpha^{-k})^n < e^{-n\alpha^{-k}} < 1 ;$$

$$\Rightarrow x_k = \alpha^k O(1)$$

$$\begin{aligned}
 \text{D'où} \quad \sum_{k < (1-\varepsilon) \log_{\alpha} n} x_k &= O(1) \frac{\alpha^{(1-\varepsilon) \log_{\alpha} (n)} - 1}{\alpha - 1} \\
 &= O(1) \frac{n^{1-\varepsilon} - 1}{\alpha - 1} \\
 &= O(n^{1-\varepsilon})
 \end{aligned}$$

$$\begin{aligned}
 Y_k &= n[e^{-n\alpha^{-k}} - (1-\alpha^{-k})^{n-1}] \\
 &= n[e^{-n\alpha^{-k}} - e^{(n-1) \log(1-\alpha^{-k})}] \\
 &= n[e^{-n\alpha^{-k}} - e^{(n-1)(-\alpha^{-k} + O(\alpha^{-2k}))}] \\
 &= ne^{-n\alpha^{-k}} [1 - e^{\alpha^{-k} - (n-1) O(\alpha^{-2k})}] \\
 &= ne^{-n\alpha^{-k}} [-\alpha^{-k} + (n-1) O(\alpha^{-2k})]
 \end{aligned}$$

En majorant $e^{-n\alpha^{-k}}$ par 1, on obtient :

$$Y_k = (n^2 \alpha^{-2k} - n\alpha^{-k}) O(1)$$

D'où

$$\begin{aligned}
 \sum_{k < (1-\varepsilon) \log_{\alpha} (n)} Y_k &= n^2 O(1) \sum_{k < (1-\varepsilon) \log_{\alpha} n} \alpha^{-2k} + n O(1) \sum_{k < (1-\varepsilon) \log_{\alpha} n} \alpha^{-k} \\
 &= n^2 O(1) \frac{\alpha^{-2(1-\varepsilon) \log_{\alpha} (n)} - 1}{\alpha^{-1} - 1} + n O(1) \frac{\alpha^{-(1-\varepsilon) \log_{\alpha} (n)} - 1}{\alpha^{-1} - 1} \\
 &= n^2 O(1) \frac{n^{-2(1-\varepsilon)} - 1}{\alpha^{-1} - 1} + n O(1) \frac{n^{-(1-\varepsilon)} - 1}{\alpha^{-1} - 1} \\
 &= O(n^{2\varepsilon}) + O(n^{\varepsilon}) \\
 &= O(n^{2\varepsilon})
 \end{aligned}$$

Nous avons montré que :

$$\overline{SOC}^{(\infty, n)} - \varphi(n) = O(n^\varepsilon) + O(n^{2\varepsilon}) + O(n^{1-\varepsilon}) + O(n^{2\varepsilon})$$

En choisissant $\varepsilon = 1/3$, il vient

$$\overline{SOC}^{(\infty, n)} - \varphi(n) = O(n^{2/3})$$

ANNEXE 2 : Calcul de la fonction φ^*

Nous avons :

$$\begin{aligned}\varphi(n) &= \sum_{k \geq 0} \alpha^k (1 - e^{-x\alpha^{-k}} - x\alpha^{-k} e^{-x\alpha^{-k}}) \\ &= \sum_{k \geq 0} \alpha^k f(x\alpha^{-k})\end{aligned}$$

où $f(n) = 1 - e^{-x} - x e^{-x}$;

Notons φ^* et f^* , les fonctions complexes définies si elles existent, par :

$$\varphi^*(s) = \int_0^\infty \varphi(x) x^{s-1} dx$$

$$f^*(s) = \int_0^\infty f(x) x^{s-1} dx$$

φ^* et f^* sont appelés les transformées de Mellin de φ et f .

f^* est définie à condition que $f(x) x^{s-1}$ soit intégrable du $]0, \infty[$;

Au voisinage de l'infini, $f(x) x^{s-1}$ équivaut à x^{s-1} qui est intégrable pour $\text{Re}(s) < 0$.

Au voisinage de 0, $f(x) x^{s-1}$ équivaut à x^{s-1} qui est intégrable pour $\text{Re}(s) > -2$.

$$\text{Notons } \Gamma(s) = \int_0^\infty e^{-x} x^{s-1} dx \quad (\text{cf [Erde]})$$

Alors pour $-2 < \text{Re}(s) < 0$

$$\begin{aligned}f^*(s) &= \int_0^\infty (1 - e^{-x} - x e^{-x}) x^{s-1} dx \\ &= \int_0^\infty (1 - e^{-x}) x^{s-1} dx - \int_0^\infty e^{-x} x^s dx\end{aligned}$$

En intégrant par partie la première intégrale on obtient

$$\begin{aligned} f^*(s) &= \left[(1-e^{-x}) \frac{x^s}{s} \right] - \int_0^\infty e^{-x} \frac{x^s}{s} dx - \int_0^\infty e^{-x} x^s dx \\ &= 0 - \frac{1}{s} \Gamma(s+1) - \Gamma(s+1) \end{aligned}$$

En utilisant la propriété de la fonction Γ :

$$\Gamma(s+1) = s\Gamma(s) ,$$

il vient :

$$\begin{aligned} f^*(s) &= -(s+1)\Gamma(s) \\ \varphi^*(s) &= \int_0^\infty \left(\sum_{k \geq 0} \alpha^k f(x\alpha^{-k}) \right) x^{s-1} dx \\ &= \sum_{k \geq 0} \alpha^k \left(\int_0^\infty f(x\alpha^{-k}) x^{s-1} dx \right) \quad \text{si cette série converge} \end{aligned}$$

En faisant le changement de variable $x\alpha^{-k} = x'$, on obtient

$$\begin{aligned} \varphi^*(s) &= \sum_{k \geq 0} \alpha^k * \alpha^{ks} \left(\int_0^\infty f(s) x^{s-1} ds \right) \\ &= f^*(s) \sum_{k \geq 0} \alpha^{k(s+1)} \end{aligned}$$

On a donc :

$$\varphi^*(s) = f^*(s) \sum_{k \geq 0} \alpha^{k(s+1)}$$

à condition que f^* soit définie et que la série $\sum_{k \geq 0} \alpha^{k(s+1)}$ converge.

Ceci est vérifié pour s $-2 < \text{Re}(s) < 0$ et $\text{Re}(s) < -1$.

Alors

$$\varphi^*(s) = f^*(s) * \frac{1}{1 - \alpha^{s+1}} = -(s+1)\Gamma(s)$$

ANNEXE 3 : Calcul de $\int_{c-i\infty}^{c+i\infty} \varphi^*(s) x^{-s} ds$;

$$\varphi^*(s) = \frac{-(s+1)\Gamma(s)}{1 - \alpha^{s+1}} \quad \text{pour } -2 < \operatorname{Re}(s) < -1 ;$$

Comme la fonction Γ, φ^* peut être prolongée par analytité et on peut l'intégrer sur le rectangle $R(N)$ défini par ses quatre sommets dans $\mathbb{C} : c-iN, d-iN, d+iN, c+iN$ ($c \in]-2, -1[$; $d \in]-1, 0[$; $N \in \mathbb{N}$)

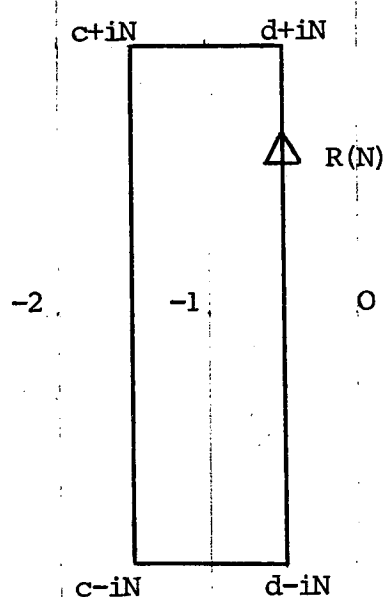


Figure : Schéma du contour $R(N)$ dans le plan complexe ;

Les pôles de la fonction $\varphi^*(s) x^{-s}$, à l'intérieur de $R(N)$ sont les racines de l'équation :

$$\begin{cases} 1 - \alpha^{s+1} = 0 \\ s \in R(N) \end{cases} \Leftrightarrow \begin{cases} s = -1 + \frac{2ik\pi}{\log \alpha} \\ k \in \mathbb{Z} \text{ et } \frac{2k\pi}{\log \alpha} \leq N \end{cases}$$

Dans $R(N)$, la fonction Γ n'a que le pôle -1 ; Notons R_k , le résidu de $\varphi(s) x^{-s}$ en $-1 + \frac{2ik\pi}{\log \alpha}$;

Au voisinage de $-1 + \frac{2ik\pi}{\log \alpha}$ ($k \neq 0$), on a :

$$\frac{1}{1 - \alpha^{s+1}} = \frac{-1}{(\log \alpha) (s+1 - \frac{2ik\pi}{\log \alpha})} + O(s+1 - \frac{2ik\pi}{\log \alpha})$$

$$\text{Donc } \varphi^*(s) x^{-s} = \frac{\frac{2ik\pi}{\log \alpha} \Gamma(-1 + \frac{2ik\pi}{\log \alpha}) x^{1 - \frac{2ik\pi}{\log \alpha}}}{(\log \alpha) (s+1 - \frac{2ik\pi}{\log \alpha})} + O(s+1 - \frac{2ik\pi}{\log \alpha})$$

$$\text{D'où } R_k = \frac{2ik\pi \Gamma(-1 + \frac{2ik\pi}{\log \alpha}) x^{1 - \frac{2ik\pi}{\log \alpha}}}{(\log \alpha)^2}$$

Au voisinage de -1 , on a :

$$\frac{1}{1 - \alpha^{s+1}} = \frac{-1}{(\log \alpha) (s+1)} + O(s+1)$$

$$\Gamma(s) = \frac{-1}{(s+1)} + (\gamma-1) + O(s+1)$$

$$x^{-s} = x - x \log x (s+1) + O(s+1)^2$$

$$\text{Donc } \varphi^*(s) x^{-s} = \frac{-x}{(\log \alpha) (s+1)} + O(1)$$

$$\text{D'où } R_0 = \frac{-x}{\log \alpha}$$

En appliquant le théorème des résidus, nous avons donc :

$$\int_{R(N)} \varphi^*(s) x^{-s} ds = 2i\pi \left[-\frac{x}{\log \alpha} + \sum_{\substack{k \in \mathbb{Z} - \{0\} \\ \frac{2k\pi}{\log \alpha} < N}} R_k \right]$$

Par ailleurs :

$$\begin{aligned}
 \int_{R(N)} \varphi^*(s) x^{-s} ds &= \int_{c-iN}^{d-iN} \varphi^*(s) x^{-s} ds \\
 &+ \int_{d-iN}^{d+iN} \varphi^*(s) x^{-s} ds \\
 &+ \int_{d+iN}^{c+iN} \varphi^*(s) x^{-s} ds \\
 &+ \int_{c+iN}^{c-iN} \varphi^*(s) x^{-s} ds \\
 &= I_1 + I_2 + I_3 + I_4
 \end{aligned}$$

Le fonction Γ vérifie la propriété :

$$\forall (x, y) \in \mathbb{R}^2 \quad \lim_{|y| \rightarrow \infty} (|\Gamma(x+iy)| e^{1/2\pi|y|} |y|^{1/2-x}) = \sqrt{2\pi}$$

(cf [Erde] Chap. 1, Vol. 1)

En particulier on en déduit que :

$$\exists K > 0 \quad \exists N_0 \in \mathbb{N} \quad \text{tq} \quad \forall N > N_0$$

$$|\Gamma(x+iN)| < K N^{x-1/2} e^{-\frac{\pi N}{2}}$$

Si x décrit $]c, d[$, alors on a :

$$|\Gamma(x+iN)| < K N^{d-1/2} e^{-\frac{\pi N}{2}}$$

De même $\exists k', \exists N_1 \in \mathbb{N} \quad \text{tq} \quad \forall N > N_1 \quad \forall x \in]c, d[$

$$|\Gamma(x-iN)| < K N^{d-1/2} e^{-\frac{\pi N}{2}}$$

On en tire que :

$$\begin{aligned}
 |I_1| &= \left| \int_{c-iN}^{d-iN} \varphi^*(s) x^{-s} ds \right| \\
 &< K N^{d-1/2} e^{-\frac{\pi N}{2}} \left| \int_{c-iN}^{d-iN} \frac{(s+1)x^{-s}}{1-\alpha^{s+1}} ds \right|
 \end{aligned}$$

La fonction $\frac{(s+1)x^{-s}}{1-\alpha^{s+1}}$ est continue donc absolument intégrable sur le compact $[c-iN, d-iN]$

On en tire que :

$$|I_1| = O(N^{d-1/2} e^{-\frac{\pi N}{2}})$$

De même on montre que :

$$|I_3| = O(N^{d-1/2} e^{-\frac{\pi N}{2}})$$

$$\begin{aligned} |I_2| &= \left| \int_{d-iN}^{d+iN} \frac{(s+1)\Gamma(s)x^{-s}}{1-\alpha^{s+1}} ds \right| \\ &\leq \int_{d-iN}^{d+iN} \frac{|x^{-s}| |s\Gamma(s)|}{|1-\alpha^{s+1}|} ds + \int_{d-iN}^{d+iN} \frac{|x^{-s}| |\Gamma(s)|}{|1-\alpha^{s+1}|} ds \end{aligned}$$

$$\begin{aligned} \text{D'où } |I_2| &\leq \frac{|x^{-d}|}{(1-\alpha^d)} \int_{-N}^{+N} (|\Gamma(d+it+1)| + |\Gamma(d+it)|) dt \\ &\leq \frac{|x^{-d}|}{(1-\alpha^d)} \int_{-\infty}^{+\infty} (|\Gamma(d+it+1)| + |\Gamma(d+it)|) dt \end{aligned}$$

Comme nous l'avons déjà vu :

$$\exists A \in \mathbb{R}^+, \exists K \in \mathbb{R}^+ / \forall t \in \mathbb{R}$$

$$\begin{aligned} |t| > A \Rightarrow |\Gamma(d+it+1)| &< K e^{-\frac{\pi|t|}{2}} |t|^{d-1/2} \\ \text{et } |\Gamma(d+it)| &< K e^{-\frac{\pi|t|}{2}} |t|^{d+1/2} \end{aligned}$$

Cela suffit à montrer que les fonction $|\Gamma(d+it+1)|$ et $|\Gamma(d+it)|$ sont intégrables en t sur les intervalles $]A, +\infty[$, $]-\infty, -A[$ (Elles sont majorées par des fonctions intégrables).

Par ailleurs, ces fonctions sont également intégrables sur le compact $[-A, +A]$, car elles sont continues.

$$\text{On en tire que } |I_2| = O(|x|^{-d})$$

Nous avons montré que :

$$\int_{R(N)} \varphi^*(s) x^{-s} ds = \int_{c+iN}^{c-iN} \varphi^*(s) x^{-s} ds + O(|x|^{-d}) \\ + O(N^{d-1/2} e^{-\frac{\pi N}{2}})$$

En prenant la limite quand N tend vers l'infini, il vient

$$\int_{c-i\infty}^{c-i\infty} \varphi^*(s) x^{-s} ds + O(|x|^{-d}) = 2i\pi \left[-\frac{x}{\log \alpha} + \sum_{k \in \mathbb{Z}-\{0\}} R_k \right]$$

D'où, en prenant $d = -1/2$:

$$\int_{c-i\infty}^{c+i\infty} \varphi^*(s) x^{-s} ds = 2i\pi \left[\frac{x}{\log \alpha} - \sum_{k \in \mathbb{Z}-\{0\}} R_k \right] + O(|x|^{1/2})$$

REMERCIEMENTS

Je tiens à remercier, pour leur aide et leurs conseils durant ce travail, Guy Fayolle et tout particulièrement Philippe Flajolet.

REFERENCES

- [Erde] Higher Transcendental Functions
Erdeleyi and Al ;
- [Flajolet, Sotteau] A recursive partitioning process on computer science.
Philippe Flajolet (INRIA), Dominique Sotteau (Univ. Paris-Sud)
Invited paper for the II World conference on "Mathematics at
the service of the man" . Las Palmas, June 1982.
- [Knuth] The art of computer programming
Vol. 3
O.E. Knuth.
- [Litwin] TRIE HASHING
W. Litwin (INRIA)
Rapport SIRIUS Nov. 1980 .
- [VERSO] A Relational Back-End Data Base Machine
Bancilhon and Al (INRIA)
International Workshop on Data Base Machine
San Diego - Sept. 82.

